指导教师：杨涛　　提交时间 20160320

CVPR2015 Paper Translation

No: 01
姓名：刘帆
学号：2013302601
班号：10011306

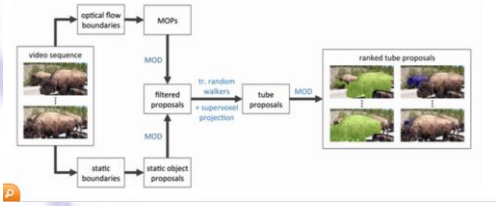
学会分割视频中的移动目标

我们从视频中分割移动目标，是按照“移动对象”特征空间段排名；他们可能包含一个移动的物体。在视频的每个帧中，计算上使用多个图地面每帧运动的边界采取分割段的特征。我们对他们进行排名与运动对象性检测有序的图像和运动领域的侦测到移动物体并丢弃上/下分割或场景的背景部分。我们使用的密集点的轨迹运动在空间中随机采取排名靠前特征。我们的最终排名始终优于在之前的视频分割方法，是目前可用的最大的视频分割基准。此外，我们对每帧移动物体的特征提取，相比以前的国家的最先进的静态建议的方法，提高检测率高达 7%。

1、介绍

建议区域可能包含对象和使用分类卷积神经网络是目前静态图像目标检测的最主要的方法。根据经验，这一地区的美国有线电视新闻网模式已经显示出性能优良的滑动窗口的分类[19]，往往不能枚举所有可能的图像边框，或 MRF 的像素分类，使独立性像素标签的组织的假设，并不能区分附近相同的对象类[9]、[16]。在本文中，我们提出了一个类似的模式，通过引入基于运动对象的特征和移动对象检测视频中的移动物体。在这项工作中，我们提出了比以前的多尺度分割和轨迹聚类方法，以及特征的生成方法，不考虑运动边界或移动物体的本质。

我们提出了一个方法，通过细分目标物体的移动视频对象生成多个分割的运动边界移动单眼未标定的视频对象，并用“移动对象性”探测器排名的方法。在每个帧中，我们通过在光流的大小施加基于学习边界检测器来提取运动边界。所提取的运动边界建立生成特征空间，这是我们每帧运动目标的特征调用分割像素，我们称之为每帧运动目标的特征空间。我们把国家的最先进的静态部分特征提取提高目标检测率和展示运动在视频对象的检测优化了 7%以上。我们每帧特征和静态的特征提取扩展到空间上使用密集点的轨迹约束分割。建议集是一个“移动对象”，从形状和光流训练的卷积神经网络检测器（MOD）探测到移动物体并丢弃和分割场景的静态部分。这个排名可以确保良好的分割对象，只有极少数的边界部分。我们的方法的概述示于图 1。



我们使用光学光流边界输入来分割，他们没有静态相结合的界限;我们通过计算分组从 RGB 和运动边缘，获得了一组不同的对象。与此相反，许多研究人员已经尝试用静态界线光流相结合，以提高边界检测[34]，[37]，得到了一定的成功[11]，[37]。这主要是由于光流对真实对象的边界失调：在“出界”跨越闭塞轮廓的背景[38]由于背景象素模仿附近前景的运

动，如图 2 所示[29]，[34]，[37]企图通过改变相邻图像区域的流动内容改变静态边界片段的强度。由静态边界检测器的性能它们的上界。对于高的阈值，许多边界被错过，没有希望得到恢复。对于低的阈值，很多的图像杂波导致的区域太小的流量进行汇总，有效地防止“出界”[37]。我们绕过出界的问题完全由直接供应稍微不一致的流动边界作为输入进行分割。

我们使用的密集点轨迹运动和随机行走来段扩展到每帧的时空管。运动是一种线索，因为作为对象不经常处于运动[10]。在帧中，它们是静态的，没有光流界限和 MOPs 错过它们。约束轨迹集群传播从有“运气”的分割，大型运动帧与很少或没有运动帧。我们再根据它们与超级像素重叠轨迹簇映射到象素管。

我们的移动对象性检测器（MOD）学习从一组训练样例移动物体的外观。它过滤了每帧段的其他爆炸性数量的建议和最后一组的排名的时空空间。移动对象的多个类表示对 RGB 和运动领域的双通道架构 CNN;其神经元捕获的鱼，人，车，动物等，并利用它们之间的相似外观，例如，许多动物有四条腿.该模型优于手工编码中心环绕显著性和其他竞争对象基线[1][20]。

我们的方法差距是在运动分割和跟踪的方法。以前的运动分词[32]，[39]操作“自下而上”，他们利用颜色或运动提示信息，而无需使用对象的训练集。以前跟踪[4]，[13]使用对象检测器（例如，汽车或行人检测器），关注场景的相关部分。我们使用一个训练集学习运动对象的概念，但仍不可知存在于视频的准确对象类。

总之，我们的贡献是：

•从光流边界对多个运动目标进行分割。

•一个移动的对象性检测器，对每帧分割特征建议排名。

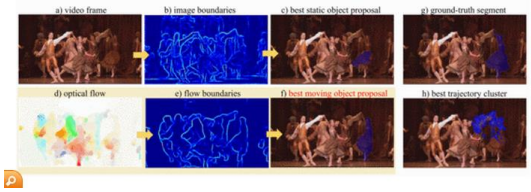
•随机走路的运动轨迹为嵌入每帧段延伸到时空轨迹集群。

我们测试的两个最大目前可用的方法的视频分割基准：moseg [6]和[12] VSB 100。我们的目标是最大限度地提高交叉口的联盟（IOU）我们的时空空间使用建议尽可能真实的物体。这是相当于标准的性能指标，在静态域产生的分部特征 [2]，[22]。在每一个视频，地面实况目标的 55%-65%使用 64-1000 管建议的具有挑战性的 VSB100 的基准被捕获，赢过[12]，[28]，[39]的竞争方法。我们的经验表明，我们的方法可以处理的铰接对象和拥挤的场景，是在对现有的方法和基线挑战下的情况。我们的代码是网站是 www.eecs.berkeley.edu/~katef/.

2、相关工作

我们可以将以前的方法分类的基础上的信息进行归类：一是自顶向下追踪法，二是自底向上的视频分割方法。跟踪方法利用特定类别的探测器把关注点放在现场的相关部分，如行人或汽车追踪器[4]，[13]。视频分割的方法无视对象的类别。作品[18]，[39]组像素的颜色和/或光流的相似性，并产生多尺度的时空象素组。每个时空的超像素被称为超体元。通过平滑的时间处理多尺度静态边界地图使用光流的工作[12]提出了国家的最先进的结果在 VSB 100 集，[6]、[28]、[32]集群密集的点的轨迹【35】使用长距离运动轨迹的相似性。他们大多对刚性物体的基准优异的显示效果。[28]的工作轨迹集群映射到一个多尺度 MRF 场的像素每帧的超像素。通过考虑高阶仿射模型建立轨迹的亲和力与轨迹稀疏。虽然[28]，[6]，[32]方法专注于获

得一个单一的轨迹聚类[6]，我们的经验发现，多尺度轨迹聚类有效地处理所造成的运动变化的对象分割歧义。我们将使用它在实验部分作为一个附加的基线。许多方法都尝试结合区域和点轨迹[28]，[30]，[33]。



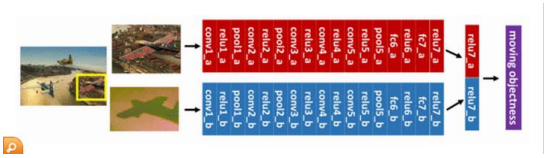
轨迹集群已显示出较大的时空视野捕捉对象比超级像素；后者是敏感的帧边界强度的波动。然而，衔接或大型运动会导致频繁的像素闭塞。轨迹太短，在这种情况下非常有用。

计算每一帧的多段建议[3]，[26]，并通过使用外观相似帧进行链接。由多个静态数字地面分割类似[25]得到的特征。工作[29]产生多个视频段通过取消图像边界，不会表现出高流动强度。两个工程都是由静态边界检测器的上限。与我们相关的是[31]的工作，从光流直接计算对象的特征。他们认为光流梯度是更嘈杂的输出的一个学习的边界检测器。此外，代替计算多个分割，他们计算出每帧的一个形象地假设，根据他们与流梯度空间关系分类像素图或地面。

许多上述方法不显示在标准基准测试结果，并与他们进行比较是困难的。在我们的实验中，我们以[12]流行的超体元方法相比，[39]和先进的轨迹聚类方法[28]，它是可扩展的，它的代码是公开的。

3、运动目标的分割(MOPs)

给定视频序列中，我们使用是 BrOx 和 Malik 的大位移光流计算每个帧中的光流场[5]。然后，我们计算光流场的边界，运用艺术结构的森林边界探测器[8]对光流场的大小，我们复制成一三个通道的图像。虽然探测器已经被训练在 BSD 系统边界基准[27]静态图像的边界，尽管不同的统计数据，但它可以有效地检测出流场的边界，如图 2 所示，例如，流量的大小有更多的圆角比人的 RGB 图像的结构。我们没有考虑重新训练检测器利用光流输入，因为与真实图像边界的流动边界的错位程度（由于流量“溢出”）广泛取决于背景，会使检测器混淆。



我们用流动边界地图诱导 KRäHeNBüHL 和 Koltun [22]用基于像素的亲和力的轮廓测对象方法。给定一个图的边界，[22]计算分割随机目标的位置并根据最短路径的目标进行超像素分类。它是最近在国家 Pascal 对象检测基准最先进的分割结果。然而，强内部图像边界会导致持续到的饱和点对象碎裂[22]，即增加的阈值没有改善目标覆盖的数目。MOPs，图 2F 所示，虽然与真实图像边界的流动“出血”的边缘没有对齐，但视频基准分割指标改善。光流有效地填充了内部的边界，由于外观或表面的标记，在运动中是平滑的，在运动中微弱的加强交叉物体的轮廓。

4、移动物体检测器

我们对运动对象检测器（MOD）使用 CNN[24]与双通路结构操作对图像和流场，如图 3 所示。对于流动通道，我们提供了一个 3 通道图像的缩放比例的位移场和光流幅度。每个网络协议栈的架构是类似于 krizhevsky 的。[23]：假定 C（K，N，S）是具有核大小为 k * K，N 滤波器和 s 的步幅的

卷积层，P（K，S）的内核大小为 k＊k 和步幅的最大值池层 S，N 正常化层，RL 整流线性单位，FC（N）与 N 个滤波器和 D，一个辍学层差比 r 一个完全连接层。每个堆栈的结构是如下：C（7，96，2） －RL－P（3,2）-NC（5，384，2） － RL － P（3,2）-NC（3，512 ，1） － RL － C（3，512,1） － RL － C（3，384,1） －RL－P（3，2） －（FC 4096） －RL-D（0.5）-FC（4096 ） －RL。图像和流量堆的特征是级联和一个最终层到输入边框的联合相交在与该地面片段。

我们初始化权重在每种网络协议栈使用 200 个对象类检测网络[15]，从 RGB 图像训练在图像检测任务。许多运动物体类别代表在图像训练集。我们也期望[15]，我们说的检测网络[23]，成为一些成立的概念对象。我们通过网络使用一个小的收集框，用来捕捉移动的物体（以及大量的背景框）的 vsbi00 和 moseg 视频基准的集合。我们使用标准的随机梯度下降用一个公开的深度学习包训练我们的 MOD。

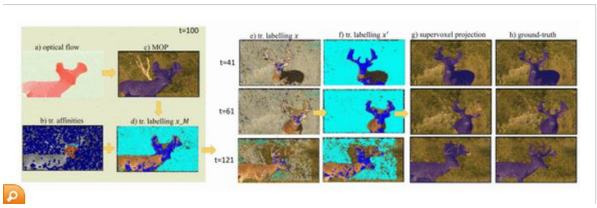
5、新的特征

我们将每帧的像素管 MOPs 的时空传播的标签通过使用随机行走的轨迹运动 [17]和轨迹簇映射到像素，如图 4 所示中。

给定一个视频序列，密集的点轨迹通过连接光流场计算 [36]。当向前向后的一致性检查失败时，表示对应的模糊轨迹终止。这通常是在像素闭塞或 DIS-闭塞的情况下，或在负荷下的图像 texturedness。让 T 表示集合的视频轨迹，让 n 表示轨迹的数目，n =T 我们计算两两轨迹亲和力 A=[0,1]n*n 区间上运动相似轨迹间是一个函数的最大速度差，由[6]提出的，因而是强大的每帧模糊的运动。我们计算每对在时间上重叠，并且是 60 象素的空间距离内的轨迹之间的亲和力。轨迹亲和性可视如图 4b。

让 ti 表示框架，MOPs 进行检测，点的轨迹相交架。被标记为前景或背景。他们分别是蓝色和淡蓝色图 4d 所示。在此之前，终止或终止之后启动轨迹是未标记。它们在图 4e 中显示。设 x={0,7}71 分别表示轨迹标签，1 表示前景和 0 表示背景。令 F 表示前景和 B 背景轨迹套，分别让 M = FNB 表示设定标记（标记）轨迹和 U= T\ M 中的一组未标记的轨迹的。让表示的轨迹的正规拉普拉斯矩阵： L = Diag（A1N）-A，其中（诊断）（y）的表示与向量（y）的对角矩阵。我们最小化在所提出的随机步行者成本函数[17]：

$$\min_{x} \; \tfrac{1}{2}x^T\mathbf{L}x \quad \text{subject to} \quad x_B = 0, \quad x_F = 1. \tag{1}$$



显而易见，减少 XTL ，x 等价于最小化。我们采取 x 真值 $\sum_{i,j}^{n} \mathrm{A}_{ij}(x_i - x_j)^9$。公式 1 有一个封闭的形式的解决方案，其中：xU 是我们正在寻找的未标记的轨迹的标签，是标签标记的轨迹。我们的近似计算的封闭形式的解决方案，通过执行一系列的标签使用归一化的基质扩散：

$$x' = \mathrm{Diag}(\mathbf{A}\mathbf{1}_n)^{-1}\mathbf{A}x. \tag{2}$$

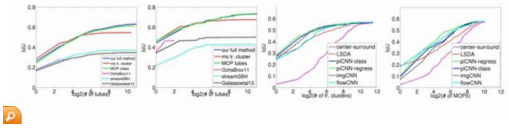
我们发现 50 的扩展是足够大约 60 像素每帧的半径。我们在图 4f 扩散轨迹标签。我们通过映射在 supervoxels，跨越多个帧扩展超像素

的加权平均轨迹集群像素。我们通过平滑时刻超像素标签，类似于[12]。每个超像素的权重是其交点超过联盟（IOU）得分与轨迹集群。我们的阈值的加权平均得到每个轨迹集群的二进制时空分割，在图 4g 所示：图中的鹿已经从它的背景被完全分割。注意，尖锐边界已尽管在图 4c 生成 MOP 的未对准边界回收。另外，通过轨迹人口稀少由于低图像 texturedness，如鹿身体图像部分，已经被正确地标记。

6、实验

我们测试我们的两个最大的公开的视频分割基准方法：VSB100[12]和 Moseg[6]。 VSB100 包含 100 个视频片段，40 个训练和 60 个测试，他们是从 Youtube 收集高清晰度视频。物体运动是非常微妙的或非常明确的。许多拥挤的场面都包括在内，如游行，自行车赛，沙滩排球，芭蕾，萨尔萨舞等等。我们专注于 VSB100 基准的刚性和非刚性运动任务是关注移动物体分割（而不是分割静态背景）。Moseg 数据集包含 59 视频，描绘从好莱坞电影“马普尔小姐”的场面，以及汽车和动物，如猫，兔子，熊，骆驼，马等移动物体有不同的运动环境和场景相对整洁，每部影片对象很少（平均一或两个）。

首先，我们完整的运动分割方法和对[28]的国家的最先进的单级点轨迹集群相比较，以及[12]的超像素方法[39]。我们的方法达到更高的地面实况报道比以往的作品为任意数量的分割。其次，我们每帧的多目标优化基准静态图像分割。我们发现，多目标优化与静态段建议结合[22]他们实现这一超越的静态部分建议的饱和点平均最佳的重叠，覆盖范围和检测率。最后，我们的基准我们对每帧段以及时空管特征排名移动对象性检测，并用其他的 CNN 结构，中心环绕显着性和静态图像对象性比较。



运动分割

我们比较[12]，[39]的流行超像素的方法和[28]的轨迹聚类和像素化方法。对于[12]，我们用自己的方法实现，因为我们超体元计算紧随他们的方法。对于[39]和[28]我们使用的代码可在网上找到。对于我们的方法，我们使用我们对移动对象的检测其时空段。得分多元化已被用于在[7]极大值抑制。分层时空分割的[18]，分布式的代码[39]，在我们的基准使用，没有足够的可扩展性。我们的国防部的战士使我们能够利用不同的管的建议。我们的 MOD 分级器允许我们利用不同的套管的特征。我们考虑多尺度轨迹聚类作为一个这样的源，即补充了我们 MOP 特征。具体地说，我们离散轨迹运动的亲和力[40]用于改变本征向量的数目的频谱嵌入。我们使用 50 如在我们所有的实验中使用的特征向量的最大数量。

我们展示在 VSB100 和 Moseg 基准图分别为 5 列 1 和 2 中，是运动分割的结果。横轴表示的每个视频序列中使用的特征数目，纵轴表示与地面实况时空段交会点，平均视频序列。我们对 MOP 评测、多尺度的轨迹集群（MS 处理集群）以及他们整体，这是我们完整的方法。我们的方法优于以前很多的提案方法。多尺度轨迹集群不提供显著提升 MOP，但对自己是一个非常有竞争力的基准。还注意到在两个数据集的所有方法的性能差异大，VSB100 超过 Moseg 更具挑战性的本质（很多非刚性物体，含蓄或阐明运动

等）的表现相差很大。

静态分割

我们在每一帧的对象分割进行 MOPs 性能测试。我们考虑以下四种广泛使用的静态图像分割指标：

一）平均最佳重叠，地面实况对象的所有段的建议平均最佳重叠在我们的数据的平均值（在所有二维的地面真实片段在我们的数据集）。二覆盖范围：权值，由地面实况段的面积加权的加权平均（大段更为重要）。三）在 50%的检测率：地面实况领域有 IOU 在以上一段方案占 50%。四）在 70%。具有在[22]该 70%的阈值要求的对象之间更知觉相似，并且是物体检测更好的度量被示出检测率。我们进一步提出任何最好的（AB）版本，C 和 D 的指标，其中每个地面管（而不是每帧段）我们认为最好的生命段的重叠；这一指标给出我们 MOP 管的性能。

我们表明，该 MOPs 的结果，[22]（GOP 的）和组合段的建议（GOP+ MOP）表 1。接下来每个方法，我们显示了括号使用数量。结合 MOPs 和 GOP 的实现了 6%的检出率分别为 50%和 70%的重叠，同比增长 5%，在充满挑战的 VSB100 标准下，并在 Moseg70%重叠检出率%增加了 5%，在相同数量的特征下。这说明的 GOP 和 MOPs 是相辅相成的，它们不能在不同的地方取得成功。性能提升是在 VSB100 数据集大。这些数字不能通过增加中，我们观察到在 2500 个数量的每帧的特征达到其饱和点[22]。

特征排名

我们测试我们的移动对象性探测器上的每帧的 MOPs 和由多尺度轨迹集群中的 VSB100 数据集聚类产生。我们展示了在第一种情况下，相应的排名曲线产生的平均在第二个在 5 和 4，分别在第 3 和第二，在第二个视频序列的图像。曲线显示多少凹陷，需要达到的与地面的特值在一特定水平的交叉口。我们定义每个管的得分的总和的边界框在其寿命值的总和。我们使用的总和，而不是平均，因为我们要对较长的特征。

我们从图像和流场（piCNN 回归）针对双通道分类 CNN（piCNN 类）比较我们的双通路 CNN 回归，图像只有美国有线电视新闻网（imgCNN），流量只有美国有线电视新闻网（flowCNN），我们的实现的从光流幅度（中央环绕）标准中心环绕显着特征度量[14]，并使用 7000 类探测器从大尺度[20]的领域适应性（LSDA）工作的 OB-jectness 探测器。我们的双通路分类 CNN 是训练来进行分类，而不是回归到他们的借条比分使用借条的 50%阈值盒为阳性或阴性。对于我们的方法基线，我们考虑 [20]检测框的置信度的每个每帧段边界框的加权平均，其中权重对应于它们的交点上与与边框。我们发现这种对象性基准有提供具有竞争力的静态对象性探测器。

我们的双通路 CNN 回归考虑的备选方案中表现最好，虽然有双通道分类 CNN 接近的性能。我们的 CNN 的网络上的节段，而不是它的分割掩码的边界框进行操作。同时掩蔽背景是可能的，背景用于判断上下分割极为重要。

讨论失败案例

在 VSB100 集，在测试过程中有很多失败的案例。它们是由大型运动或完全遮挡物体造成的。我们的方法以及我们的基线将受益于额外的连接步骤，其中，类似地寻找管挂横跨以形成较长的。为了保持清洁的方法，我们没有考虑这样的一个步骤。在镶嵌数据集，最失败的案例是由于像素管

轨迹集群的不准确映射：我们经常会稍微漏到背景，尤其是对四肢比较细的动物，如骆驼。

时间计算

下面的数字是一个 CPU 值。大排量光流平均 16 秒/张。给定一个光流场，计算多目标优化需要 4 秒钟一个 700×1000 的图像。每个 MOPs 轨迹嵌入的投影需要 2 秒 70000 的轨迹，都可以多目标优化同时使用矩阵运算扩散投射。超体素计算是因果和需要 7 秒每帧。计算运动亲和力 70000 轨迹中的每个视频需要 15 秒。我们的超像素计算，光流计算，MOP 计算和预测是完全并行。

7、总结

我们已经提出了细分影片由多段分割生成根据移动的对象性运动的物体和排名的方法。我们的运动物体分割保证他们捕捉移动物体的表现补充静态的，特别是在混乱，具有挑战性的场景。所提出的移动物体检测器丢弃上方和下方分割或场景的背景部分，并提供了一个排名，允许在每个视频捕获几个地面真实的对象。该方法填补了视频分割和跟踪研究的差距，通过利用训练集学习移动物体的外观。

致谢

我们衷心感谢为本研究 Tesla GPU 的捐赠的 NVIDIA 公司。